Assessment of model uncertainty for soil moisture through ensemble verification

Gabriëlle J. M. De Lannoy,[1] Paul R. Houser,[2] Valentijn R. N. Pauwels,[1] and Niko E. C. Verhoest[1]

Received 14 June 2005; revised 14 October 2005; accepted 25 January 2006; published 17 May 2006.

[1] The Community Land Model (CLM2.0) has been used to simulate land surface processes in a small corn field. The subdivision of grid cells into patches in the CLM2.0 was explored for the generation of Monte Carlo simulations for use in calibration and ensemble generation. A distributed multiobjective calibration was developed for the optimal estimation of parameters and initial state variables for 36 soil moisture profiles. Since the resulting parameter and initial state values did not lead to perfect simulations for soil moisture, and in order to better understand the forecast uncertainty, ensemble runs were generated. The ensembles generated by CLM2.0 have been verified by several methods that are commonly used in meteorology. It was shown that the perfect model approach cannot be applied for bounded hydrological applications and that perturbation of parameters is a necessity to obtain a realistic assessment of the forecast error. Perturbation of forcings only captures more of the model uncertainty than perturbation of initial conditions only, but also causes a too limited spread in the ensembles. The generation of ensemble members through perturbation of the parameter set, found through calibration, does not necessarily result in ensembles that surround the calibrated deterministic control run for soil moisture. This is partially due the nonlinearity of the model in the parameters. It may also indicate that some parameter sets are not robust and not appropriate to perturb for ensemble generation. Consequently, the resulting ensemble mean may not represent the best forecast or a priori state estimation. During periods of extreme drought or precipitation, the ensemble probability density function (pdf) deviates far from normality and the model behaves very nonlinearly. For state estimation, methods like the ensemble Kalman filter are best suited for the propagation of the first moments to account for the nonlinear dynamics during crucial events for hydrological simulations. However, the a posteriori estimate for this technique will only be optimal in the limited class of linear filters, since the underlying pdfs cannot be assumed to be Gaussian.

Citation: De Lannoy, G. J. M., P. R. Houser, V. R. N. Pauwels, and N. E. C. Verhoest (2006), Assessment of model uncertainty for soil moisture through ensemble verification, *J. Geophys. Res.*, *111*, D10101, doi:10.1029/2005JD006367.

1. Introduction

[2] Reproducing the observed behavior of a system by a model requires a proper system identification. For hydrological applications, a wide variety of physical models are known, in which relationships between physical variables in a natural system are mapped onto mathematical structures. Once the structure of a model is (assumed to be) known, the parameters should be optimized through parametric identification methods, which is referred to as calibration.

[3] While the majority of calibration studies have been concerned with lumped applications, more and more research on the calibration of distributed models has been reported [*Boyle et al.*, 2001; *Houser et al.*, 2001], mostly, however, after a drastic initial reduction in the number of parameters, obtained by keeping several parameters constant in space. Efficient and effective fully distributed calibration is still a topic of research, to which this paper attempts to contribute.

[4] Given a model structure, calibration is preceded by the selection of parameters to optimize [*Bastidas et al.*, 1999], the optimization method [*Boyle et al.*, 2000], the objective functions [*Gupta et al.*, 1998] and the calibration period [*Yapo et al.*, 1996]. Also, the choices of the initial state [*Gao et al.*, 1996] and input forcings [*Xia et al.*, 2005] affect the calibration. It should be emphasized that calibration leads to optimal parameter values that are conditioned

[1] Laboratory of Hydrology and Water Management, Ghent University, Ghent, Belgium.
[2] George Mason University and Center for Research on Environment and Water, Calverton, Maryland, USA.

0148-0227/06/2005JD006367$09.00

on these choices, the model structure (and error) and the observational data error.

[5] Even when a model is well calibrated, some uncertainty in model forecasts will always remain. Because of this model uncertainty, *Krzysztofowicz* [2001] stressed that forecasts should always be stated in probabilistic, rather than in deterministic, terms. Best estimates of a state should always be accompanied by a quantification of uncertainty. One possibility to obtain this information is to generate an ensemble of realizations.

[6] The first applications of ensemble forecasting for weather prediction only tried to assess the uncertainty in forecasts due to the growth of errors in the initial conditions. This approach is called the strong constraint or perfect model approach. The uncertainty in a priori estimates/forecasts is not only due to errors in the initial conditions, but also to model error, for example, due to a simplified model structure or to unsatisfactory parameterization. Inclusion of the simulation of model error therefore provides a more realistic idea of the spread in forecasts. This is called the weak constraint or imperfect model approach. While methods for the generation of ensemble members have received a lot of attention in meteorology and oceanography, they hardly have been considered in hydrology. This may partially be explained by the nonchaotic nature of hydrological models, and consequently the relatively small impact of very small deviations from the optimal initial state. Furthermore, hydrological studies use time-dependent external forcings that influence the evolution of the ensemble forecasts. Such problems are called boundary value or boundary-forced problems. The possibility of providing ensembles in the field of hydrology has only recently emerged [*Butts et al.*, 2004; *Georgakakos et al.*, 2004; *Carpenter and Georgakakos*, 2004; *Lee and Anagnostou*, 2004; *Krzysztofowicz*, 2001] and has mostly been explored in the field of ensemble data assimilation for state estimation [*Reichle et al.*, 2002a, 2002b; *Margulis et al.*, 2002; *Crow and Wood*, 2003].

[7] Ensemble forecasting typically generates an overwhelming amount of information that is difficult to analyze in detail. *Stephenson and Doblas-Reyes* [2000] described and applied several statistical methods for interpreting Monte Carlo ensemble forecasts for multivariate systems. *Toth et al.* [2003] summarized methods of verification for probabilistic forecasts of scalar variables, and in particular for ensemble forecasts developed for meteorological purposes. In this paper, several of these methods will be applied.

[8] In this study, field-scale soil moisture processes have been simulated by a land surface model. In sections 2 and 3 the data and model are described and some model adaptations are discussed for the generation of ensemble runs. Calibration of the model using soil moisture data in order to obtain optimal parameter and initial state estimates is discussed in section 4. As these parameter estimates will not render a perfect model, the a priori estimates/forecasts by the model will be uncertain. To assess this uncertainty, the generation and verification of ensembles is described in sections 5 and 6. A proper verification is important to better understand the behavior of ensemble statistics extracted from the ensemble probability density functions (pdf). Ensemble realizations and their statistics are often considered to represent the true uncertainty around the truth. However, this assumption is only valid if good ensembles are generated. While verification of ensembles has received considerable attention in meteorology, in hydrology ensembles tend to be generated in a relatively simple manner and used without further investigation. In this paper, the performance and characteristics of ensemble runs for soil moisture are studied using techniques that are commonly used in meteorology. Conclusions are drawn in section 7.

2. Data Description

[9] The Optimizing Production Inputs for Economic and Environmental Enhancement (OPE^3, http://hydrolab.arsusda.gov/ope3/) project is an interdisciplinary research project which was started in 1998 and is managed by the Beltsville Agricultural Research Center (BARC)–Agricultural Research Service (ARS) of the United States Department of Agriculture (USDA). The project is conducted on a corn field of 21 ha, subdivided into four fields. The site is situated in Prince Georges County, Maryland, and it is part of the Anacostia watershed. The four subwatersheds are named A, B, C and D from north to south. Water draining from the field feeds a wooded riparian wetland and first-order stream, Beaver Dam Creek, which subsequently drains into the Anacostia River, the Potomac River and the Chesapeake Bay.

2.1. Soil Moisture Observations

[10] In each subwatershed of the OPE^3 field, 12 capacitance probes (EnviroSCAN, SENTEK Pty Ltd., South Australia) have been installed to measure volumetric water contents within a 10 cm radius from the sensor's center [*Starr and Paltineanu*, 2002]. The sampling interval is 10 min. To compare the data to the model results, the observations were aggregated into hourly time steps.

[11] The probes are named following a three-digit system. The first letter represents the name of the subwatershed (A, B, C, D), the second letter (L, H, M) refers to the estimated infiltration rate at the point of installation (Low, High, Moderate) and the third digit (1, 2, 3, 4) discerns between the different probes of a specific infiltration regime. H-probes have sensors at 10, 30 and 80 cm. L- and M-probes have sensors at 10, 30, 50, 120, 150 and 180 cm. L-probes have an additional sensor at 80 cm depth.

[12] In this study, data collected from 1 May 2001 through 30 April 2002 were used. During this period, probes AL3, AL4, AM3, AM4, AH3, AH4, CL3, CL4, CM3, CM4, CH3 and CH4 were not operational owing to lightning damage, reducing the operational number of probes to 36.

2.2. Atmospheric Forcings

[13] The meteorological data required for the Community Land Model (CLM2.0), which will be described in the next section, are air temperature (K), wind speed (m/s), specific humidity (kg/kg), incident solar radiation (W/m^2) and total precipitation (mm/s). Other forcings are calculated by the model itself.

[14] In field B of the OPE^3 corn field, meteorological data are collected at a 10-min interval by instruments on the 10-m-high USDA meteorological tower. Data from 9 June

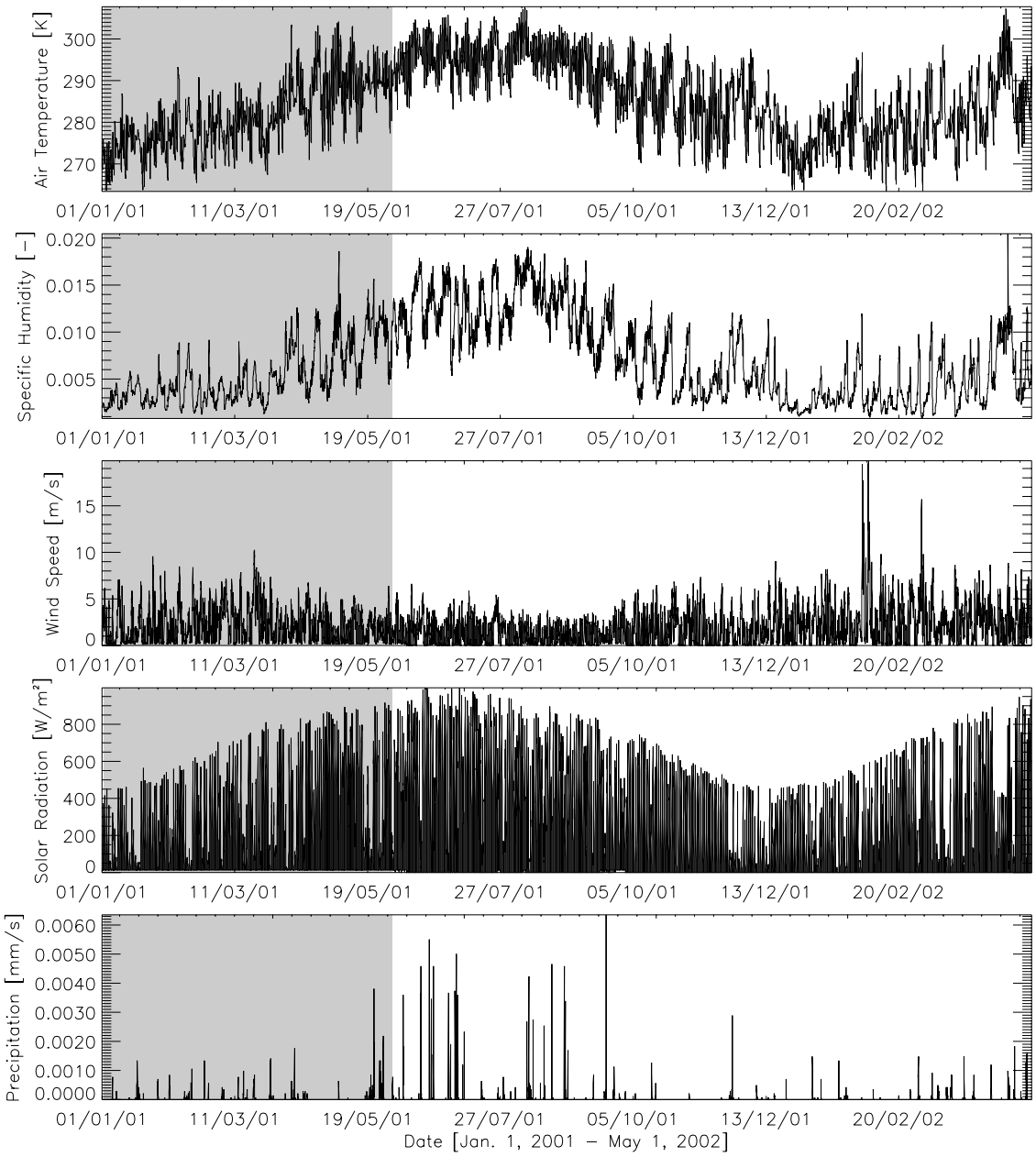

Figure 1. Meteorological data used to force CLM2.0. The data in the gray background are purely based on data from the Station 3 Old Beltsville Airport. Meteorological data in the B field are available after this period only.

2001 through 30 September 2002 were made available by the BARC-ARS of the USDA Hydrology and Remote Sensing Lab. These data cover the most important variables to force hydrological models (air temperature, relative humidity, wind speed, solar radiation, and precipitation). To deal with missing data, data from two towers outside the field were used. A meteorological tower of the Soil Climate Analysis Network (SCAN) is situated just outside field D. The height of the tower is approximately 3.5 m. The vegetation cover at the site is grass. Data from this tower are recorded hourly and are available through the Internet from October 2001 to present. The Station 3 Old Beltsville Airport is another site that is situated relatively close to the OPE^3 field. Data at a 15-min interval from January through December 2001 were provided by the BARC-ARS USDA Farm Operations Branch. The site has a 3.05-m-high tower. The data from the tower in field B were mainly used, and missing data were filled in through regression with the data from the other towers. The resulting time series of forcings are shown in Figure 1.

[15] The observed atmospheric forcings were assumed to be spatially uniform. This assumption is made, even though it is well known that the spatial variability of precipitation is the main influence on model output, and that lack of spatial information in the calibration procedure causes the optimal parameter sets to compensate for these errors in input.

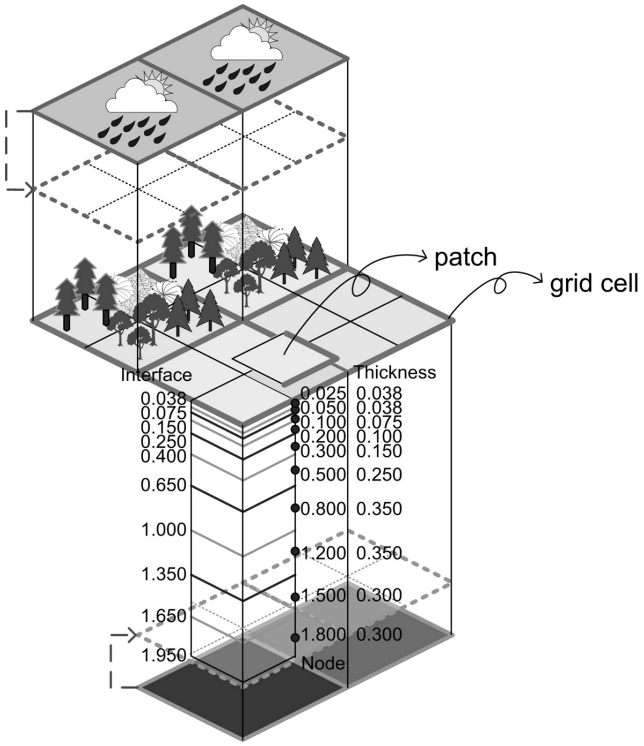

Figure 2. CLM2.0 structure. The dashed lines represent the changes to the original structure: Soil parameterization and forcings are assigned to patches instead of grid cells. The depth and thickness of each soil layer are given in meters.

However, since the study area is smaller than the size of a typical precipitation cell [*De Lannoy et al.*, 2005], the spatial variability of rainfall in this field is small.

3. Model Description

[16] The Community Land Model (CLM) is a land surface model for which the initial CLM code was completed in 1998 by combining the best features of three existing modular land models: the Biosphere-Atmosphere Transfer Scheme (BATS) [*Dickinson et al.*, 1993], the Land Surface Model (LSM) [*Bonan*, 1996], and the model developed at the Institute of Atmospheric Physics (IAP94) [*Dai and Zeng*, 1996] in Beijing. *Zeng* [2003] reported some recent experiences from this CLM project. In this work, CLM2.0 was used in offline mode, without any coupling to an atmospheric or climate model.

[17] The CLM2.0 models biogeophysical and other processes over a predefined grid by calculating water and heat fluxes and states for every grid cell separately, without any interaction between cells. Each grid cell can be subdivided into several patches, containing one single land cover type: vegetation, bare soil, wetland, lake, urban and glacier. In this study, each grid cell was completely covered with vegetation. The vegetated fraction is further subdivided into patches of plant functional types. Each patch maintains its own prognostic variables. By default, all patches within a grid cell have the same (grid cell) soil texture, soil color, and corresponding physical properties and they respond to the same mean conditions (forcings) of the overlying atmospheric grid cell.

[18] For the model applications in this study, it was desirable that the patches differed in more characteristics than only in land cover. The possibility to change (perturb) the mean atmospheric forcings for every single patch was developed. Furthermore, the possibility to attribute different soil characteristics to each patch was introduced. Figure 2 gives a schematic overview of the model structure, including the adaptations for this study.

[19] CLM2.0 has one vegetation layer, a user-defined number (by default 10) of vertical soil layers, and up to five snow layers (depending on the snow depth). For this work, the choice for the depths of the nodes of the different soil layers was based on the depths of the soil moisture observations and the need for thin surface layers to assure numerical convergence. The depths of the different soil nodes were set to 2.5, 5, 10, 20, 30, 50, 80, 120, 150 and 180 cm depth. The corresponding layer thickness and the depths of the interfaces are given in Figure 2. It is clear that the difference in layer thickness will be a source of representativeness error for soil moisture when model results are compared to observations. The model was integrated forward with a constant hourly time step.

4. Calibration of Parameters and Initial State Estimation

4.1. Global Optimization Method

[20] A purely random, Monte Carlo (MC) search was performed to calibrate the CLM2.0 for soil moisture through estimation of parameters and initial state variables. Basically, this approach transformed the classical calibration problem into a weak-constraint variational data assimilation problem, in which an optimal estimate of the initial state and parameters was sought. The CLM2.0 was run forward for each patch or MC member with different parameter vectors (a parameter vector contains all model parameters and the initial state variables), and each corresponding time series of soil moisture was compared to observations through an objective or cost function. The best parameter vector was the one with the lowest value for the objective function. Because of the large number of parameters and their complicated interactions, a limited sensitivity analysis did not allow for a proper selection of parameters, without the risk of assigning badly defined constant values to parameters that would not be included in the calibration. Therefore it was decided to calibrate with brute force, perturbing all parameters without any effort to reduce the dimension of the parameter space. Of course, the problem is that each additional parameter increases the minimum attainable uncertainty on the individual parameter estimates [*Cramér*, 1946].

[21] Traditional MC simulations take parameters from a uniform distribution. However, in this research, the parameters were drawn from Gaussian distributions, which were truncated, in order to not include physically or numerically impossible parameter values. A Gaussian distribution was chosen, as for most parameters the exact distribution representing the uncertainty on the parameters is not known, and it can be expected that the shape of the distributions is different for each parameter. We found from experiments

Table 1. Surface Data Used in CLM2.0[a]

Surface Data	μ_C	σ_C	m	s	min	max
Soil color index	1,2,3,4,5,6,7,8		4.611	2.348	1	8
Percentage sand at 2.5 cm, %	62.17	5.56	V	V	V	V
Percentage clay at 2.5 cm, %	15.62	1.63	V	V	V	V
Monthly averaged LAI, Jan	0.1	0.01	0.102	0.01	0.078	0.122
Monthly averaged LAI, Feb	0.1	0.01	0.103	0.011	0.081	0.124
Monthly averaged LAI, March	0.1	0.01	0.098	0.01	0.078	0.121
Monthly averaged LAI, April	0.2	0.01	0.197	0.009	0.175	0.219
Monthly averaged LAI, May	0.5	0.1	0.476	0.082	0.314	0.688
Monthly averaged LAI, June	1.5	0.5	1.586	0.54	0.694	2.731
Monthly averaged LAI, July	3.5	0.5	3.462	0.464	2.52	4.333
Monthly averaged LAI, Aug	4	0.5	3.922	0.505	3.046	4.867
Monthly averaged LAI, Sept	3.5	0.5	3.448	0.471	2.307	4.324
Monthly averaged LAI, Oct	0.5	0.1	0.49	0.108	0.321	0.757
Monthly averaged LAI, Nov	0.1	0.01	0.102	0.011	0.07	0.121
Monthly averaged LAI, Dec	0.1	0.01	0.101	0.011	0.078	0.123
Monthly averaged top height, Jan, m	0.01	0	0.01	...	0.01	0.01
Monthly averaged top height, Feb, m	0.01	0	0.01	...	0.01	0.01
Monthly averaged top height, March, m	0.01	0	0.01	...	0.01	0.01
Monthly averaged top height, April, m	0.01	0	0.01	...	0.01	0.01
Monthly averaged top height, May, m	0.13	0.05	0.129	0.047	0.018	0.255
Monthly averaged top height, June, m	0.85	0.16	0.847	0.121	0.604	1.109
Monthly averaged top height, July, m	2.06	0.3	2.026	0.295	1.292	2.541
Monthly averaged top height, Aug, m	2.2	0.3	2.187	0.32	1.42	2.805
Monthly averaged top height, Sept, m	2.15	0.3	2.178	0.309	1.54	2.861
Monthly averaged top height, Oct, m	0.01	0	0.01	...	0.01	0.01
Monthly averaged top height, Nov, m	0.01	0	0.01	...	0.01	0.01
Monthly averaged top height, Dec, m	0.01	0	0.01	...	0.01	0.01

[a]Here μ_C and σ_C determine the distribution for the generation of MC runs for calibration, and m, s, min, and max represent the mean, standard deviation, minimum, and maximum values of the obtained best parameters over all 36 calibrated soil profiles. V denotes that texture varies with depth and is perturbed to calibrate soil physical variables only (see Table 4).

using uniform distributions for the parameter perturbation that more resulting parameter combinations were invalid for use in the model, since there was a higher risk of combining (sometimes contrasting) extreme values for different parameters. Further, the intention was to generate zero mean Gaussian model state errors through Gaussian ensemble perturbations. However, we recognize that since the model is nonlinear, the resulting state errors may not be Gaussian, which would require additional study. The default parameter values were kept as mean values. Through Gaussian perturbation, a higher probability for an appropriate mean value results in a higher probability to obtain physically realistic parameters, or parameters that are effective for the model. Through trial and error by studying numerical and physical problems in the model results, the standard deviation for perturbation was chosen to be maximally of the same order of magnitude as the mean value (see Tables 1, 2, 3, 4 and 5). Some boundary limits were imposed to avoid impossible values, such as, for example, negative values for the hydraulic conductivity. When a parameter value beyond some predefined limiting bounds was generated, a new value was drawn.

[22] Instead of multiple running (restarting) the CLM2.0 at identically the same grid cell, the subdivision of grids and patches was used to calculate MC realizations by simulating over all patches in space. A rectangular grid of 10×10 cells was chosen over the area bounded by the outer boundaries of the OPE^3 field, with each grid cell containing 1500

Table 2. As in Table 1 but for Parameters Related to the Vegetation

Physiology of Plant Functional Types	μ_C	σ_C	m	s	min	max
Momentum roughness length to canopy top height	0.12	0.05	0.110	0.047	0.038	0.224
Displacement height to canopy top height	0.68	0.05	0.696	0.048	0.585	0.778
Characteristic leaf dimension, m	0.04	0.01	0.040	0.011	0.020	0.062
Photosynthetic pathway: 0 = C4, 1 = C3	0	0	0	0	0	0
Max carboxylation at 25°C, μmol CO_2 m^{-2} s^{-1}	50	2	49.858	2.303	46.169	53.946
Slope conductance-to-photosynthesis relationship	9	0.5	9.101	0.555	8.400	10.424
Quantum efficiency at 25°C, μmol CO_2 μmol photon^{-1}	0.06	0.05	0.055	0.041	0.000	0.147
Leaf reflectance, visible	0.11	0.05	0.096	0.047	0.006	0.203
Leaf reflectance, near infrared	0.58	0.05	0.572	0.042	0.486	0.660
Stem reflectance, visible	0.36	0.05	0.369	0.044	0.289	0.475
Stem reflectance, near infrared	0.58	0.05	0.551	0.045	0.461	0.650
Leaf transmittance, visible	0.07	0.05	0.084	0.047	0.005	0.231
Leaf transmittance, near infrared	0.25	0.05	0.267	0.045	0.168	0.331
Stem transmittance, visible	0.22	0.05	0.217	0.044	0.133	0.284
Stem transmittance, near infrared	0.38	0.05	0.360	0.044	0.248	0.424
Leaf/stem orientation index	−0.3	0.1	−0.292	0.095	−0.466	−0.108
Rooting distribution parameter a, m^{-1}	6	0.05	6.005	0.049	5.892	6.097
Rooting distribution parameter b, m^{-1}	3	0.05	2.995	0.043	2.905	3.105

Table 3. As in Table 1 but for Time-Invariant Physical Constants

Time-Invariant Physical Constants	μ_C	σ_C	m	s	min	max
Roughness length for soil, m	0.01	0.01	0.014	0.008	0.003	0.028
Roughness length for snow, m	0.0024	0.001	0.003	0.001	0.000	0.004
Drag coefficient for soil under canopy	0.004	0.001	0.004	0.001	0.003	0.006
Maximum dew, mm	0.1	0.1	0.131	0.087	0.002	0.291
Fraction of model area with high water table	0.3	0.1	0.264	0.082	0.097	0.463
Tuning factor for temperature	0.34	0.1	0.347	0.085	0.174	0.528
Crank Nicholson factor between 0 and 1	0.5	0.01	0.497	0.009	0.471	0.515
Irreducible water saturation of snow	0.033	0.01	0.032	0.010	0.015	0.056
Limit of porosity before impermeability	0.05	0.01	0.048	0.012	0.023	0.067
Ponding depth, mm	10	5	10.520	4.137	2.394	17.327
Wilting point potential, mm	−1.5E+5	1.0E+5	−1.57E+5	0.89E+5	−4.03E+5	−0.06E+5
Restriction for minimal soil potential, mm	−1.0E+8	1.0E+8	−1.13E+8	0.66E+8	−2.95E+8	−0.02E+8
Water table depth scale parameter, m^{-1}	1	1	1.340	0.773	0.065	2.854
Saturated soil hydraulic conductivity bottom, mm s^{-1}	4.0E-2	0.01	0.036	0.011	0.014	0.059
Base flow parameter for saturated fraction, mm s^{-1}	1.0E-5	1.0E-5	1.3E-5	0.8E-5	0.00	2.7E-5
First bottom layer contributing to the base flow	5	1	6.056	0.911	4	8
Last top layer contributing to the surface runoff	3	1	3.222	0.946	2	5

patches. Through this combination of grid cells and patches, $15 \cdot 10^4$ MC simulations were generated, and from this collection of simulations, a best parameter set was extracted for each sensor.

4.2. Calibration Period and Objective Functions

[23] The model runs were initiated on 1 January 2001. A calibration period of 1 month was chosen in September 2001 (from 2 September 2001 through 1 October 2001). In this period, observations showed no evidence of lateral flow, as could be observed for some preceding months. Including this phenomenon would result in parameters that try to compensate for structural model errors, since the model does not simulate horizontal water flow. As optimal parameter estimation depends on the choice of the initial conditions, an initial state estimation was included in the calibration. The state variables during the 24 hours on 3 May 2001 were chosen to include the initial conditions, and their optimal values were obtained during calibration. The remaining part of the observational data was used to study the model performance in predictive mode (validation).

[24] A multiobjective calibration procedure was developed, considering different measures of goodness-of-fit, and different time series of soil moisture at the different depths in a profile. The misfit between the modeled initial state during the day of 3 May 2001, and the observations was penalized 2 orders of magnitude more (factor 100) than the misfit during September, to mimic the common practice of using observations as best initial guess for the initial conditions. This results in a least squares objective given by

$$RMSE_{ic} = \sqrt{\frac{1}{N+Nic}\left\{100\sum_{j=a}^{a+Nic}\left(y_j - x_j\right)^2 + \sum_{i=b}^{b+N}\left(y_i - x_i\right)^2\right\}}, \quad (1)$$

Table 4. As in Table 1 but for Time-Invariant Physical Constants at Two (of the Ten) Soil Layers for the Soil Moisture/Temperature Profile

Time-Invariant Physical Constants	μ_C	σ_C	m	s	min	max
			j = 1 at 2.5 cm			
b [*Clapp and Hornberger*, 1978]	5.4	1	3.375	0.746	1.633	5.128
Volumetric soil water at saturation (porosity)	0.41	0.1	0.407	0.083	0.230	0.579
Saturated hydraulic conductivity at surface, mm s^{-1}	0.008	0.1	0.083	0.067	0.001	0.277
Hydraulic conductivity at saturation, mm s^{-1}	0.0011	0.1	0.111	0.086	0.006	0.329
Minimum soil suction, mm	116.3	50	109.924	45.447	17.578	238.202
Thermal conductivity, W m^{-1} K^{-1}	7.6	1	8.919	1.034	6.297	10.819
Bulk density, mg cm^{-3}	1591	50	1615.784	234.685	1138.067	2157.023
Thermal conductivity, soil minerals, W m^{-1} K^{-1}	3.3	1	3.623	1.281	1.971	6.063
Thermal conductivity, saturated soil, W m^{-1} K^{-1}	2.6	1	3.243	1.421	0.510	5.799
Thermal conductivity, dry soil, W m^{-1} K^{-1}	0.23	0.1	0.251	0.141	0.023	0.552
Heat capacity, soil solids, J m^{-3} K^{-1}	2179604	1.0E+6	2135265	866870	358465	4267264
			j = 5 at 30 cm			
b [*Clapp and Hornberger*, 1978]	5.4	1	3.746	0.810	2.256	6.628
Volumetric soil water at saturation (porosity)	0.41	0.1	0.475	0.069	0.330	0.601
Saturated hydraulic conductivity at surface, mm s^{-1}	0.008	0.1	0.071	0.053	0.002	0.185
Hydraulic conductivity at saturation, mm s^{-1}	0.0011	0.1	0.068	0.052	0.001	0.185
Minimum soil suction, mm	116.3	50	153.753	33.245	101.031	202.545
Thermal conductivity, W m^{-1} K^{-1}	7.6	1	8.931	1.247	6.371	11.731
Bulk density, mg cm^{-3}	1591	50	1412.135	189.630	1057.555	1772.265
Thermal conductivity, soil minerals, W m^{-1} K^{-1}	3.3	1	2.792	0.867	1.131	4.783
Thermal conductivity, saturated soil, W m^{-1} K^{-1}	2.6	1	2.243	1.050	0.260	5.262
Thermal conductivity, dry soil, W m^{-1} K^{-1}	0.23	0.1	0.233	0.113	0.036	0.483
Heat capacity, soil solids, J m^{-3} K^{-1}	2179604	1.0E+6	2410856	1082396	656651	5036797

Table 5. As in Table 1 but for Initial State Variables

Initial State Variables (Start 1 Jan 2001)	μ_C	σ_C	m	s	min	max
Initial vegetation temperature, K	275	5	277.498	3.049	273.618	283.634
Initial soil-snow temperature $j = 1$, K	275	5	278.309	3.578	273.196	286.388
Initial soil-snow temperature $j = 5$, K	275	5	277.642	3.603	273.204	289.187
Initial water in canopy	0	0.1	0.065	0.043	0.008	0.208
Initial soil moisture $j = 1$	0.3	0.1	0.284	0.076	0.076	0.427
Initial soil moisture $j = 2$	0.3	0.1	0.282	0.088	0.031	0.435
Initial soil moisture $j = 3$	0.3	0.1	0.283	0.089	0.059	0.417
Initial soil moisture $j = 4$	0.3	0.1	0.280	0.087	0.129	0.449
Initial soil moisture $j = 5$	0.3	0.1	0.282	0.095	0.039	0.448
Initial soil moisture $j = 6$	0.3	0.1	0.326	0.089	0.081	0.515
Initial soil moisture $j = 7$	0.3	0.1	0.277	0.096	0.077	0.510
Initial soil moisture $j = 8$	0.3	0.1	0.253	0.078	0.085	0.406
Initial soil moisture $j = 9$	0.3	0.1	0.281	0.102	0.075	0.481
Initial soil moisture $j = 10$	0.3	0.1	0.286	0.066	0.117	0.457

with x the modeled state variables, y the corresponding observations and the subscript ic referring to the inclusion of initial conditions. The hourly time steps are denoted as $i \in [b, b+N]$ and $j \in [a, a+Nic]$, with $N = 24 \cdot 30 = 720$ for the 1-month calibration period and $Nic = 24$ for 1 day of initial conditions. Time step b is the first time step on 2 September 2001 and time step a is the first hour on 3 May 2001. Since it is known that different measures of goodness-of-fit result in different optimal parameter sets, parameter combinations, which produce good model results for multiple objective functions, were sought. Therefore, additionally, the Root Mean Square Error (*RMSE*), Nash-Suttcliffe criterium (*NS*), correlation (*R*), and absolute mean difference (*BIAS*) were calculated over the 1-month calibration period in September for each MC simulation to check the temporal evolution of soil moisture,

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=b}^{b+N}(y_i - x_i)^2} \tag{2}$$

$$NS = 1 - \frac{\sum_{i=b}^{b+N}(y_i - x_i)^2}{\sum_{i=b}^{b+N}(y_i - \langle y\rangle)^2} \tag{3}$$

$$R = \frac{\sum_{i=b}^{b+N}(y_i - \langle y\rangle)(x_i - \langle x\rangle)}{\sqrt{\sum_{i=b}^{b+N}(y_i - \langle y\rangle)^2 \sum_{i=b}^{b+N}(x_i - \langle x\rangle)^2}} \tag{4}$$

$$BIAS = |\frac{1}{N}\sum_{i=b}^{b+N} y_i - \frac{1}{N}\sum_{i=b}^{b+N} x_i| = |\langle y\rangle - \langle x\rangle| \tag{5}$$

where bracketed notation refers to temporally averaged variables.

[25] For most probes in the field, different layers of observations can serve for the calibration of one soil column. Measures of goodness-of-fit for the different layers were aggregated to one measure through simple averaging, so that for each probe or land column one value for a measure of goodness-of-fit was available. In order to obtain a combined measure over all objective functions for each land column, aggregation of the different types of measures of goodness-of-fit was performed by computing the Euclidean distance, D, of the positions (d) simulations take relative to the best simulation after sorting (best to worst) on different criteria. For example, the third best simulation for *RMSE* gets a d_{RMSE} value of 3. The Euclidean distance, D, for each land column is given by

$$D = \sqrt{d_{RMSE_{ic}}^2 + d_{RMSE}^2 + d_{NS}^2 + d_R^2 + d_{BIAS}^2}. \tag{6}$$

4.3. Multiobjective Calibration

[26] Iterative sorting of the values for the different objective functions was performed and after each sort the worst patches were excluded for further competition. After a first sorting on the Euclidean distance D, half of the patches were excluded. As the selection proceeded for the individual measures of goodness-of-fit, less patches were excluded each time the sorting was performed. The sorting procedure passed twice through a series of selection criteria, with following numbers indicating the percentage (subjective choice) remaining patches after sorting on the following sequence of criteria: Euclidean (combined) distance D: 50%, $RMSE_{ic}$: 50%, *NS*: 50%, *R*: 50%, *BIAS*: 70%, *RMSE*: 70%, *BIAS*: 50%. Thus, starting from $15 \cdot 10^4$ patches, the last sorting algorithm (second pass through the sequence of criteria, ending with *BIAS*) worked on 35 patches only. Clearly, there was a subjective choice to limit the bias and a different sorting procedure would result in a different optimal parameter set, which refers to the equifinality concept [*Beven*, 1993; *Beven and Freer*, 2001], and the concept of multiobjective equivalence of [*Gupta et al.*, 1998]. However, comparison of how parameters sets were sorted by different individual objectives revealed that some agreement can be found in the selection of the best patches for different criteria, in particular for the *BIAS* and *RMSE*. The correlation coefficient R was found to sort patches in a different way.

[27] In Tables 1, 2, 3, 4, and 5, the mean, standard deviation, minimum and maximum values of each parameter for all 36 calibrated profiles are summarized. It is clear that a large spread on the parameters is found for the different soil profiles, even though they are all situated within a relatively small area.

[28] In Table 6, the resulting averaged measures of goodness-of-fit for some selected probes are summarized. Remark that these aggregated measures are calculated for

Table 6. Calibration and Validation Measures of Goodness-of-Fit

	Calibration						Validation			
Sensor	$RMSE_{ic}$, vol%	*RMSE*, vol%	*BIAS*, vol%	*NS*	*R*	*D*	*RMSE*, vol%	*BIAS*, vol%	*NS*	*R*
BH1	5.73	1.73	0.68	0.40	0.98	1226.07	2.91	2.05	−2.37	0.87
BH2	6.92	1.55	0.50	0.18	0.98	3062.42	2.91	1.58	−0.76	0.87
BH3	5.84	1.38	1.13	−0.71	0.95	4959.59	2.16	1.28	−0.35	0.89
BH4	9.11	1.36	0.92	−0.63	0.96	3390.33	3.21	1.56	−0.14	0.86
BL1	20.40	4.88	3.74	−11867.83	0.77	6288.15	9.03	6.65	−117.03	0.47
BL2	19.56	4.46	3.61	−2158.60	0.79	2373.84	8.81	5.07	−6.13	0.55
BL3	17.05	4.21	2.81	−1028.43	0.95	6027.98	6.60	4.94	−232.98	0.69
BL4	12.87	5.57	4.53	−1775.54	0.95	4802.80	6.06	4.37	−132.96	0.89
BM1	7.24	3.94	2.72	−5838.98	0.79	5061.42	4.46	3.21	−71.67	0.86
BM2	18.76	7.95	6.15	−56364.58	0.47	1118.64	9.01	7.36	−1450.91	0.55
BM3	9.60	5.10	3.20	−4509.13	0.96	13035.31	4.75	3.15	−75.52	0.63
BM4	17.29	6.74	5.08	−154.29	0.97	1750.56	6.55	4.77	−190.74	0.87

different numbers of layers in a profile, depending on the available observations for each probe. The values reveal that calibration using only three depths of data (H-profiles) is usually easier than calibration for six or seven depths. It is logical that it is more difficult to find an optimal parameter set that yields good model results for many soil layers than for only a few. Also, the deeper layers show evidence of preferential flow, which cannot be captured by the model. However, the results for the model depths for which no observations were available for calibration, cannot be checked, and may therefore deviate from the truth. The model performance for a complete profile which was apparently successfully calibrated for a few layers, may therefore be worse than for a complete profile which was calibrated for many layers.

4.4. Validation

[29] To evaluate the model performance for predictions with the resulting optimal parameter sets, a validation period was chosen to start on 3 October 2001. The complete remaining part (split sample) of the observed soil moisture data set was used for validation, i.e., until 1 May 2002.

[30] The same (aggregated) measures of goodness-of-fit as for the calibration were used (except for the $RMSE_{ic}$ and the combined Euclidean distance *D*) and are summarized in Table 6 for some selected probes. Of course, also for validation, the values become worse when soil moisture is used for calibration at more depths. The entire validation runs with optimal parameters for the different soil profiles will be referred to as the control runs in the remainder of this paper.

[31] Note that the *NS* values suggest overall bad performances and a better performance for the validation period than for the calibration period. The explanation is that for soil moisture in deeper layers, the temporal variability in the observations during the one month of calibration is minimal and hence it causes very large negative values of *NS* for these layers. The averaged *NS* over the whole profile (i.e., for the different layers of observations) is highly influenced by these extreme negative values. The validation spans a longer time period and hence for the deeper layers the variability in soil moisture observations is larger and consequently the *NS* for these deeper layers is less negative.

5. Ensemble Generation

[32] CLM2.0 ensembles were generated by exploring the strong as well as the weak constraint approach. Different types of ensemble runs were performed: (1) perturbing initial states only, (2) perturbing parameters only, (3) perturbing forcings only, (4) perturbing parameters and initial states, and (5) perturbing parameters, initial states and forcings. Perturbation of only the initial states is interesting in order to follow up the effect of predictability error only and to study the strong constraint approach in land surface applications. From a limited sensitivity study, we found that the largest portion of the a priori estimation/forecast error was caused by occasionally badly specified parameters resulting from calibration over a short time period. Another source of uncertainty were the forcings. The quality of the forcing data may be assumed to be quite good in general. Nevertheless, differences in rainfall were found for the different meteorological stations, and the lack of spatial variability in all forcings may also contribute to some forecast error.

[33] Even though these three sources of uncertainty may capture a large part of the uncertainty in the forecast models, it is very likely that some other sources may contribute significantly, such as the uncertainty in the model physics and representativeness error. We found, for instance, that different definitions of the discretization of the soil layers resulted in slightly different model outputs.

[34] Perturbation of the optimally estimated (through calibration) initial state variables on 3 May 2001 would result in a period with unbalanced state variables during the study period of interest (when observations are available), as the perturbations of the several state variables are independent. Since we were not interested in the transient behavior, it was traced back which initial condition on 1 January 2001 resulted in the corresponding estimated (through calibration) initial condition on 3 May 2001, and the model was started from perturbation of this initial condition on 1 January 2001. The disadvantage of this method is that there is a possibility that, through the perturbation, the ensemble mean will deviate from the control value by the time it reaches the date of 3 May 2001 and that consequently the model has not been run starting from perturbation of the optimal initial conditions on 3 May 2001, nor with the corresponding optimal parameters.

[35] Generation of ensemble members was performed by perturbation of parameters and initial states around the optimal mean found by calibration and around the observed values for the forcings. Again, a similar Gaussian perturba-

tion as used for the calibration was applied, with a standard deviation chosen to be a fraction of the standard deviation used for the generation of MC realizations for calibration. Again, for perturbed values exceeding some predefined limiting bounds, new values were drawn. For this application, a grid cell was assigned to each profile and the patches in it were used for the generation of ensemble members.

[36] For the surface data, soil texture was not explicitly perturbed, as it is only used to determine soil parameters, which were directly perturbed. As soil color is assigned to grid cells (and not to individual patches), soil color was kept constant and equal to the optimal value found through calibration. The monthly leaf area index and monthly height of the top of the vegetation were perturbed around the optimal mean, found through calibration, and with a standard deviation of 1/2 $\times$ the standard deviation used in the calibration procedure. For the perturbation of the plant physiological parameters, 1/4 $\times$ the standard deviation used in the calibration procedure was applied. All time invariant physical constants were perturbed with a standard deviation of 1/2 $\times$ the standard deviation used in the calibration procedure. The initial state variables were also perturbed with a standard deviation of 1/2 $\times$ the standard deviation used in the calibration procedure. To study the impact of perturbations on the initial state only, the perturbations and limits were kept the same as for the calibration, but the standard deviation for perturbation on the initial soil moisture was increased to 50 vol%. Forcing data were perturbed around their measured values with a standard deviation of 1 K for temperature, 0.01 m/s for wind speed, 1.10^{-4} kg/kg for specific humidity, 5.10^{-5} mm/s for precipitation, 5 W/m^2 for downward longwave radiation and 50 Pa for surface pressure. Zero precipitation was not perturbed and for all forcings only strictly positive values were allowed. This may sometimes lead to a slight overestimation of the mean forcings.

6. Ensemble Verification

6.1. Ensemble Interpretation

[37] For the interpretation of the ensembles, some moments of the pdfs were studied for soil moisture. The first four moments of a pdf at time instant i are estimated by the ensemble mean, $\bar{\hat{x}}_i$, the ensemble spread, $ensp_i$, the skewness, $skew_i$, and the kurtosis, $kurt_i$, respectively,

$$\bar{\hat{x}}_i = \frac{1}{N}\sum_{k=1}^{N} \hat{x}_{i,k} \tag{7}$$

$$ensp_i = \frac{1}{N}\sum_{k=1}^{N} \left(\hat{x}_{i,k} - \bar{\hat{x}}_i\right)^2 \tag{8}$$

$$skew_i = \frac{1}{N}\sum_{k=1}^{N} \left[\frac{\hat{x}_{i,k} - \bar{\hat{x}}_i}{\sqrt{ensp_i}}\right]^3 \tag{9}$$

$$kurt_i = \frac{1}{N}\sum_{k=1}^{N} \left[\frac{\hat{x}_{i,k} - \bar{\hat{x}}_i}{\sqrt{ensp_i}}\right]^4 - 3 \tag{10}$$

with N the number of members in an ensemble, and $\hat{x}_{i,k}$ the kth member of an ensemble. It is assumed that large ensemble sizes are used, so that the division by N (instead of $N-1$) does not result in a biased estimate for the higher moments.

[38] If in the equation for the spread (equation (8)), $\bar{\hat{x}}_i$ is replaced by an observation y_i, then the mean squared error mse_i of all the ensemble forecasts is found,

$$mse_i = \frac{1}{N}\sum_{k=1}^{N} \left(\hat{x}_{i,k} - y_i\right)^2 \tag{11}$$

$$= ensp_i + ensk_I, \tag{12}$$

with $ensk_i$ the ensemble skill given by

$$ensk_i = \left(\bar{\hat{x}}_i - y_i\right)^2. \tag{13}$$

If the verifying observation is statistically indistinguishable from the ensemble members, then $ensp_i = mse_i$. In general, and for chaotic models in particular, the value of mse_i is larger than $ensk_i$, because of dispersion of ensemble forecasts $\hat{x}_{i,k}$.

6.2. Ensemble Verification Measures

[39] Some verification methods for ensembles, commonly used in meteorology to measure reliability or consistency, were used to assess the quality of the ensembles for soil moisture. *Talagrand et al.* [1997] defined several spread-skill relationships, as measures of the degree of statistical consistency between the a priori predicted uncertainty and the a posteriori or observed error in the forecast. For example, it is expected that on average the ensemble mean differs from the observation by a value that is equal to the time average of the ensemble spread. Therefore $\frac{\langle ensk\rangle}{\langle ensp\rangle}$ should reach 1, with $\langle\rangle$ meant as the average over available observations (in time). Larger values indicate too small a spread, if the model is not biased.

[40] Another measure is the ratio of the time-averaged *RMSE* of the ensemble mean to the time-averaged *RMSE* of the individual ensemble members, $\frac{\langle\sqrt{ensk}\rangle}{\langle\sqrt{mse}\rangle}$, which should equal $\sqrt{(N+1)/2N}$ [*Brankovic et al.*, 1990] if the truth is statistically indistinguishable from a member of the forecast or analysis ensemble [*Anderson and Anderson*, 1999].

[41] To predict the forecast uncertainty of continuous scalar variables, *Talagrand et al.* [1997] and *Anderson* [1996] introduced histograms of the position of the a posteriori or observed verification with respect to the a priori predicted ensemble values over some predefined period. These histograms are called rank histograms, binning diagrams, or Talagrand diagrams. For a consistent ensemble, the truth should fall into each bin with equal probability. If the histogram is flat, reliability is usually concluded, and a U-shaped histogram indicates lack of variability in the ensemble, while the opposite is true for an n-shaped histogram. An L- or J-shaped histogram corresponds to moist or dry bias (i.e., too high or low soil moisture values), respectively, for the model. The diagrams give information additionally to the measures defined

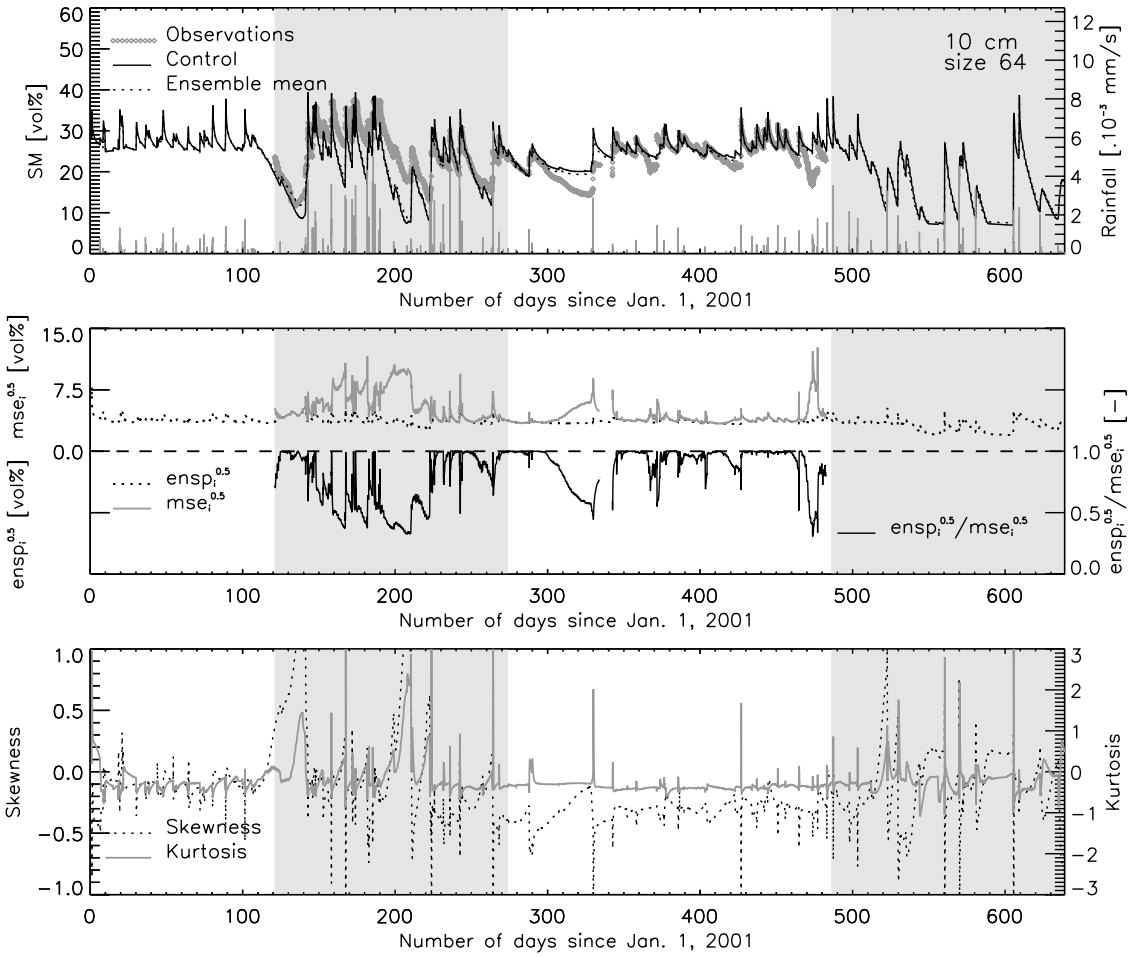

Figure 3. Evolution of some statistics for sensor BH1 at 10 cm for 64 ensemble members obtained by perturbation of initial states as well as parameters. The growing season is indicated by the gray background.

above. For example, if $\frac{\langle ensk \rangle}{\langle ensp \rangle} > 1$, the shape of the Talagrand diagram allows to discern between the presence of bias or a too small spread. *Hamill* [2001] explored the correct interpretation of rank histograms and warned for possible misleading conclusions drawn from the shape of the diagrams.

6.3. Spaghetti Plots, Histograms, and Moments: Time Series

[42] The evolution of the ensemble pdf in time can be presented in different forms. Time series of model results obtained by the different members in an ensemble can be plotted all together and overlaid by the observations, the control run, and the ensemble mean. At each time step, the information in such spaghetti plots can be summarized in a histogram. It was found for most profiles in the OPE^3 field that the histograms stayed unimodal, the ensemble member runs were highly correlated, and the control run as well as most observations remained within the range of the ensemble distributions, when the model was well calibrated.

[43] Analysis of the ensemble moments in time gives further information on the evolution of the distribution's shape in time. As an example, time series of the ensemble mean (and observations and control run for reference), standard deviation, skewness, and kurtosis are shown in Figure 3 for sensor BH1 at 10 cm depth and in Figure 4 for sensor BH1 at 30 cm depth for an ensemble size of 64 members, for perturbation of initial states and parameters. Figures 5 and 6 show some statistics for the perturbation of forcings only and for the perturbation of initial states only, respectively.

6.3.1. Spread

[44] When only initial states are excessively perturbed, the spread is so small that the control run as well as most observations are only rarely located within the range of the distributions. The spread is considerable for some initial time steps only and decreases quickly: It does not represent the uncertainty of the forecast. Perturbation of forcings only also causes a very limited spread. Parameter perturbation causes a diversity in the states from the first time step on, which is maintained during the model run time, without the need for explicit initial state perturbation. Independent of the choice of perturbation, the spread does not increase in time as, for example, atmospheric models and the different member runs show more or less the same dynamics.

[45] At each rainfall event, the spread increases, even when forcing data are not perturbed. This may be explained by the different reaction of each patch to rainfall, as each

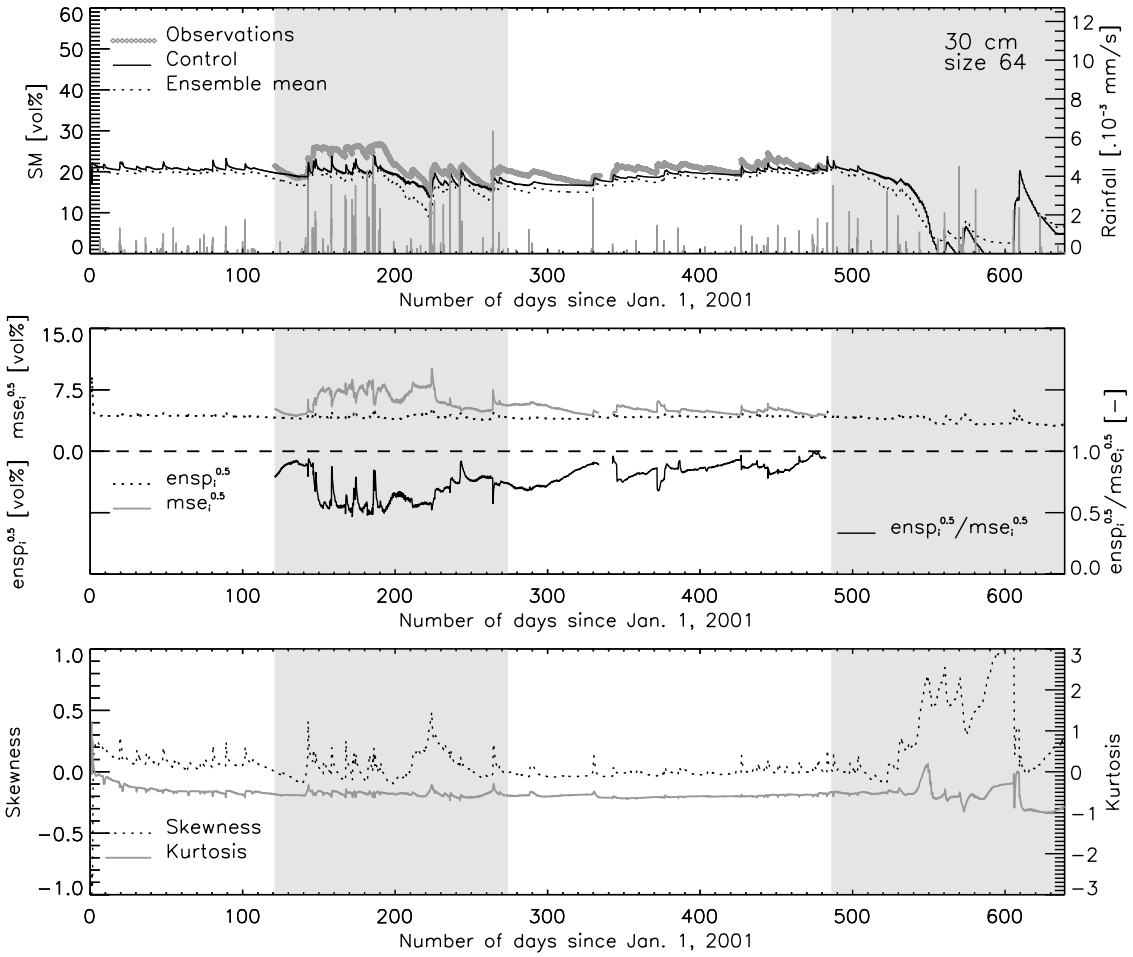

Figure 4. Evolution of some statistics for sensor BH1 at 30 cm depth for 64 members obtained by perturbation of initial states as well as parameters. The growing season is indicated by the gray background.

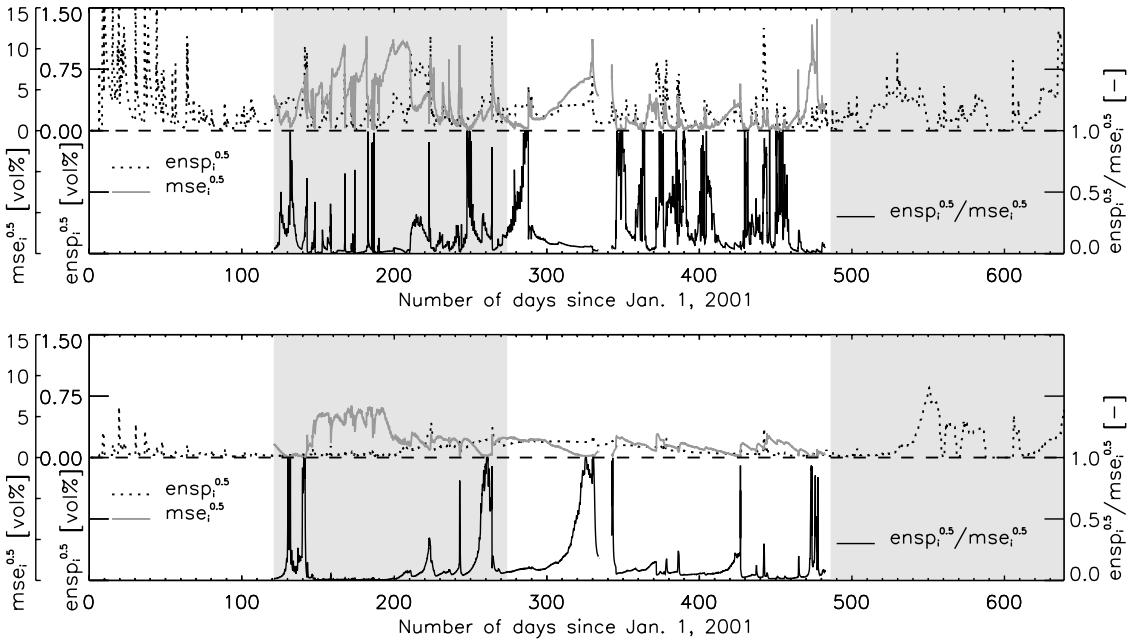

Figure 5. Evolution of $\sqrt{ensp_i}$ and $\sqrt{mse_i}$ for sensor BH1 (top) at 10 cm and (bottom) at 30 cm depth for 64 ensemble members generated by perturbation of forcings only. The growing season is indicated by the gray background.

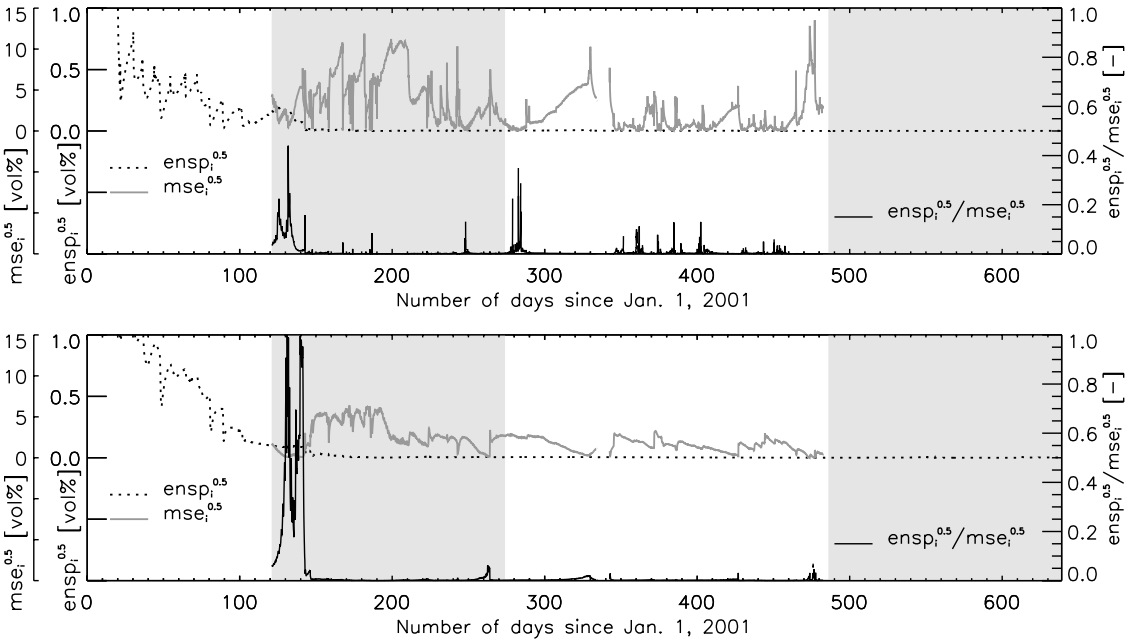

Figure 6. Evolution of $\sqrt{ensp_i}$ and $\sqrt{mse_i}$ for sensor BH1 (top) at 10 cm and (bottom) at 30 cm depth for 64 members generated by extreme perturbation of the initial states only. Note that the scale on the right vertical axis differs from the one in the previous plots. The growing season is indicated by the gray background.

patch is characterized by different parameters. A remarkable phenomenon that was sometimes observed for time series of the spread generated by perturbation of initial variables only for many members is that the different ensembles basically become identical after some time, but that after a year, a stressed moment (drought) causes a renewed dispersion in the ensembles (e.g., for sensor BH1 around day 610, in Figure 7 for 512 members). Even though the soil moisture state for all members becomes (very close to) equal, they evolve differently after the stress disappears, because the diagnostic state variables in some surrounding layers still show variability within the different members.

6.3.2. Mean Squared Error and Spread

[46] From Figures 3 and 4 for probe BH1 and for all probes in general, it is clear that $\sqrt{mse_i}$ and $\sqrt{ensp_i}$ have the same magnitude when initial states as well as parameters are perturbed, which is reassuring. However, they do not show a same evolution in time, and $\sqrt{mse_i}$ becomes much larger

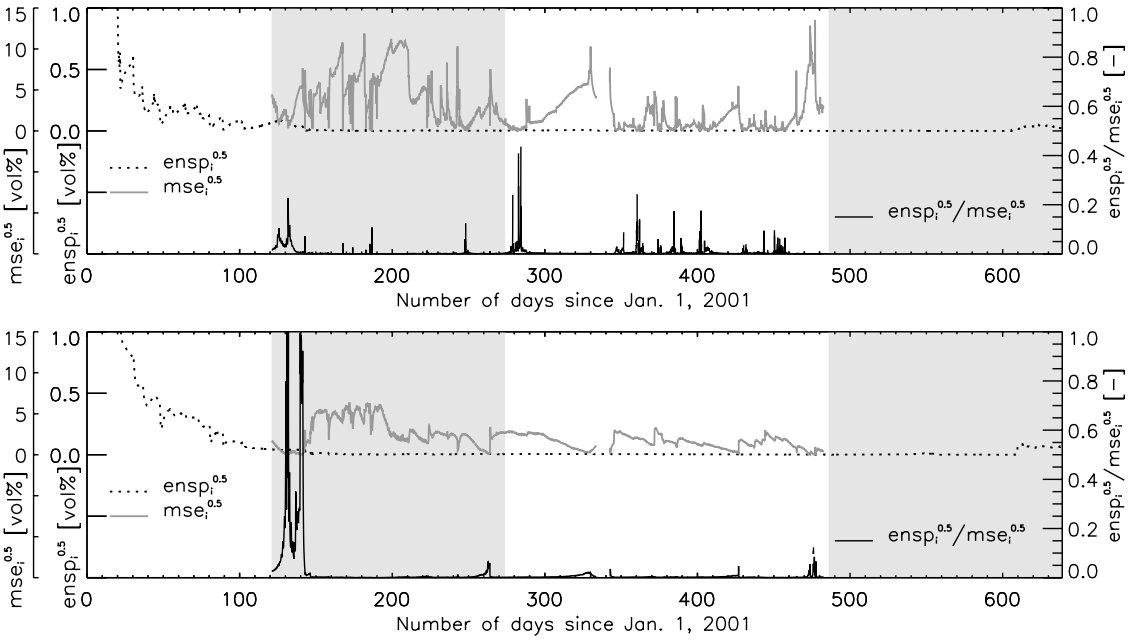

Figure 7. Same as Figure 6 but for 512 members.

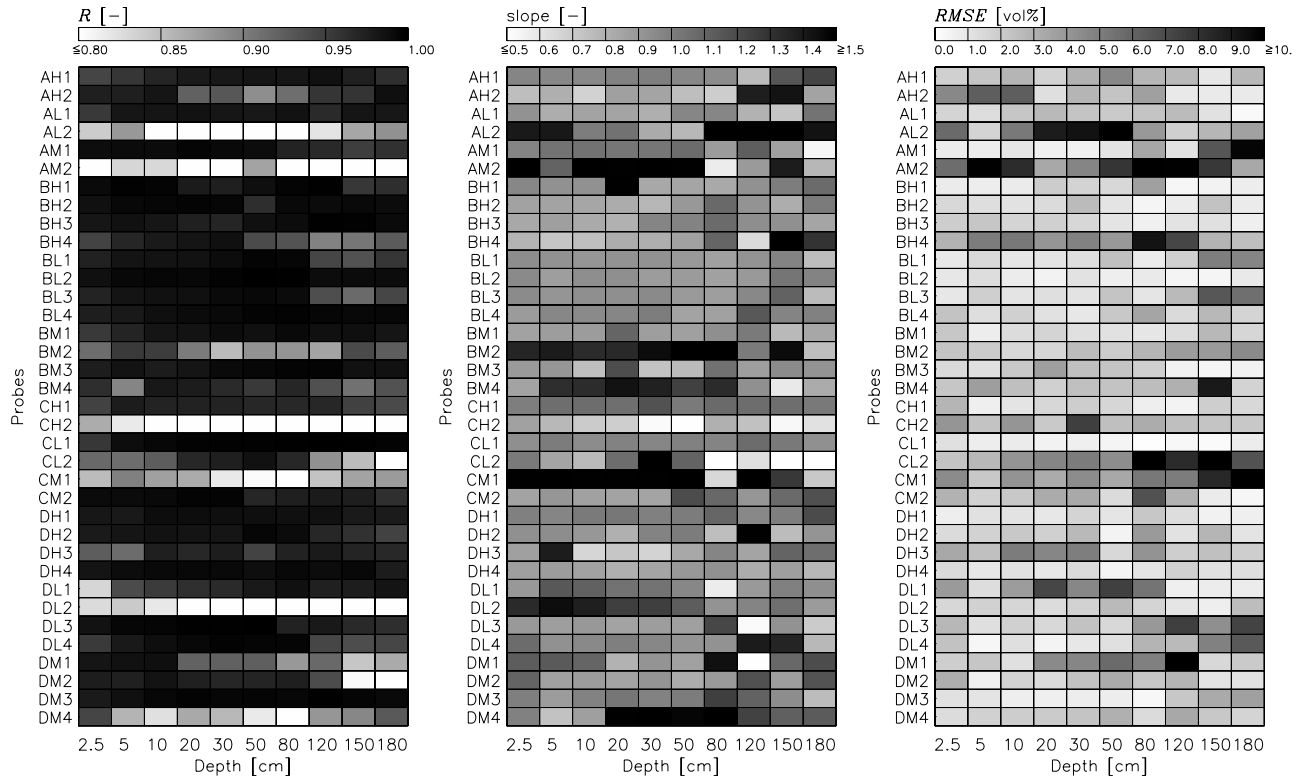

Figure 8. Correlation coefficient (R), slope and $RMSE$ between time series of the control run and the ensemble mean run for 64 ensemble members, generated by perturbation of initial states and parameters, at all model depths for all sensors.

when the model output deviates further from the observations. Decrease of the ratio $\sqrt{ensp_i}/\sqrt{mse_i}$ means a decrease of the contribution of the state dispersion to the total error and thus an increase of $ensk_i$. Figure 6 shows that the values of $\sqrt{ensp_i}$ and $\sqrt{mse_i}$ differ strongly when only initial states are perturbed: The spread induced by perturbation of initial states only is far too small to cover the range of observations, even while an extreme spread on the initial conditions for soil moisture was imposed. Clearly, $\sqrt{ensp_i}$ caused by perturbation of forcings only (Figure 5), often explains more of the $\sqrt{mse_i}$, but still this spread is too limited to describe the forecast uncertainty.

6.3.3. Skewness and Kurtosis

[47] The skewness and the kurtosis (Figures 3 and 4) remain quite close to 0 during the winter, but deviate during the growing season, when parameters and initial conditions are perturbed. During the entire period and most clearly for the upper soil layers (Figure 3), rainfall causes peaks (up and down) in skewness and kurtosis, indicating that Gaussian pdfs do not represent well the state's distributions during such events, and that strongly nonlinear processes occur. For smaller ensemble sizes, the skewness and kurtosis are even more deviating from 0. In the summer of 2002, some very dry periods stressed the corn field. The model captures this phenomenon and forces all ensembles to result in dry soil with a considerable part reaching a limiting minimum value, which causes the distribution to deviate far from Gaussian temporarily, as can be seen in the values of the skewness and the kurtosis. When only the initial states are perturbed, the skewness and kurtosis show rapid changes in values through time and deviate strongly from 0 (data not shown). This is obvious, since the initial state perturbation does not result in a well-shaped distribution, but rather in a constant value shortly after the initial time steps. This again is an argument to reject the idea of initial state perturbation only. Also, for perturbation of forcings only, the variability in skewness and kurtosis is very high (data not shown), but less than for perturbation of initial states only.

6.3.4. Remarks

[48] Because we found that the model behaves very nonlinearly during periods of precipitation or drought stress, linearization of land surface models for some state estimation techniques like the Kalman filter [*Kalman*, 1960], or the extended version of this method, will inevitably lead to inaccurate a priori estimates during these crucial periods. Alternatives like the ensemble Kalman filter [*Evensen*, 1994; *Reichle et al.*, 2002b], circumvent this problem. However, the state update equation in the ensemble Kalman filter yields optimal (minimum variance) a posteriori state estimates only if Gaussian pdfs are assumed, while in our study and other studies on land surface processes [*Reichle et al.*, 2002a; *Crow*, 2003] often non-Gaussian pdfs were found. If the pdf for the a priori state estimate is non-Gaussian, the a posteriori estimate is optimal in the class of linear filters only and better estimates may be expected from nonlinear filters [*Miller et al.*, 1999].

6.4. Time Averages of Ensemble Statistics and Verification Scores

[49] It is not straightforward that Gaussian perturbation of, for example, the optimal parameters results in a Gaussian distribution of the state around a best value for the

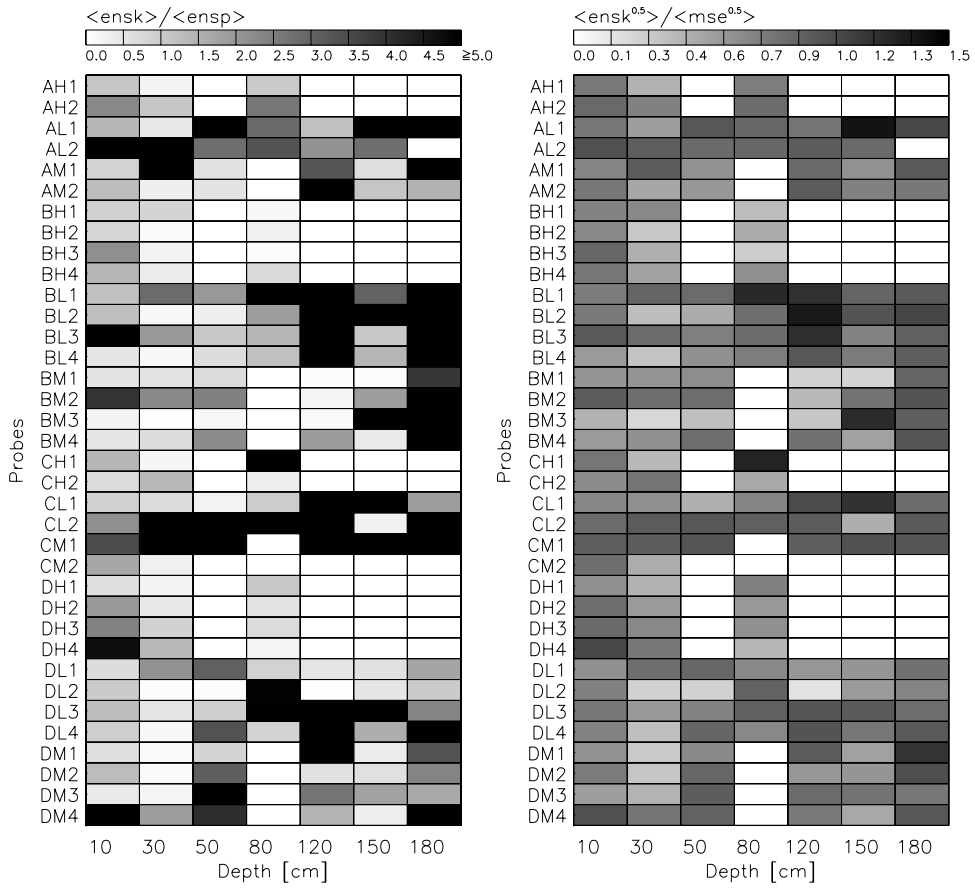

Figure 9. Skill measures at all observation depths for all working sensors for 64 ensemble members generated by perturbing initial states and parameters. White is for missing sensors.

control run. Because of the limited ensemble size and nonlinearities, the ensemble mean will inevitably differ slightly from the control run. To check if the assumption of zero mean error was valid for the ensemble runs, the correlation and slope of the regression between the control run and the ensemble mean was determined for all modeling depths (see Figure 8 for perturbation of parameters and initial states). For all temporally integrated scores, the 1-year period from 1 May 2001 to 1 May 2002 was considered. The slope of the regression line is always close to 1 and the correlation coefficient is always very high. However, sometimes there is a bias between the ensemble mean and the control run, which can be deduced from the values of the *RMSE*. For instance, for probe AL2, it was found that, starting from an initially identical value, the ensemble mean and the control for some depths take a very distinct value after a week, and an almost constant difference persists during the model period. For profile depths with high *RMSE* values, the Gaussian perturbation of the optimal parameters does not result in a Gaussian ensemble pdf around the control run. This is logical, since the model is nonlinear in the parameters. It may also be explained by the fact that the response surface of the objective function (as used in the calibration) is not smooth and that some optimal parameter sets found were situated at isolated maxima/minima, which may limit the predictive capability of the model. Such a parameter set makes a model less robust, and perturbation results in parameter sets that are far from optimal. Despite these findings, most studies on ensemble data assimilation (e.g., ensemble Kalman filter) apply random perturbation of parameters, without taking into account a possible bias effect due to perturbation of nonrobust parameter sets and the nonlinearity of the model in the parameters. However, for most profiles in the OPE[3] field, and for any ensemble size, it can be concluded though that through the perturbation of the parameters (and initial conditions) a zero mean model error is imposed.

[50] To evaluate the quality of the ensembles generated by perturbation of initial conditions and parameters, Figure 9 shows that the values for $\frac{\langle ensk \rangle}{\langle ensp \rangle}$ and $\frac{\langle \sqrt{ensk} \rangle}{\langle \sqrt{mse} \rangle}$ indicate good statistical consistency for some profiles, and bad values for others. Values of $\frac{\langle ensk \rangle}{\langle ensp \rangle}$ are around 1 for most probes, referring to corresponding ensemble spread and misfit between model output and observations. At some depths for some sensors, this ratio clearly exceeds 1, while the ratio $\frac{\langle \sqrt{ensk} \rangle}{\langle \sqrt{mse} \rangle}$ is around 1 (e.g., for probe BL2 and BL4 at 120 cm and 180 cm depth). This is caused by bias or a too small spread, which can be determined by the Talagrand diagrams (see Figure 10).

[51] Talagrand or rank histograms were generated over 1 year (1 May 2001 to 1 May 2002) for perturbation of initial states and parameters, and shown in Figure 10 for a fixed size of 64 members for all sensors in field B at all observation depths. Several plots show that the spread is too small to capture the variability in the observations. Since for

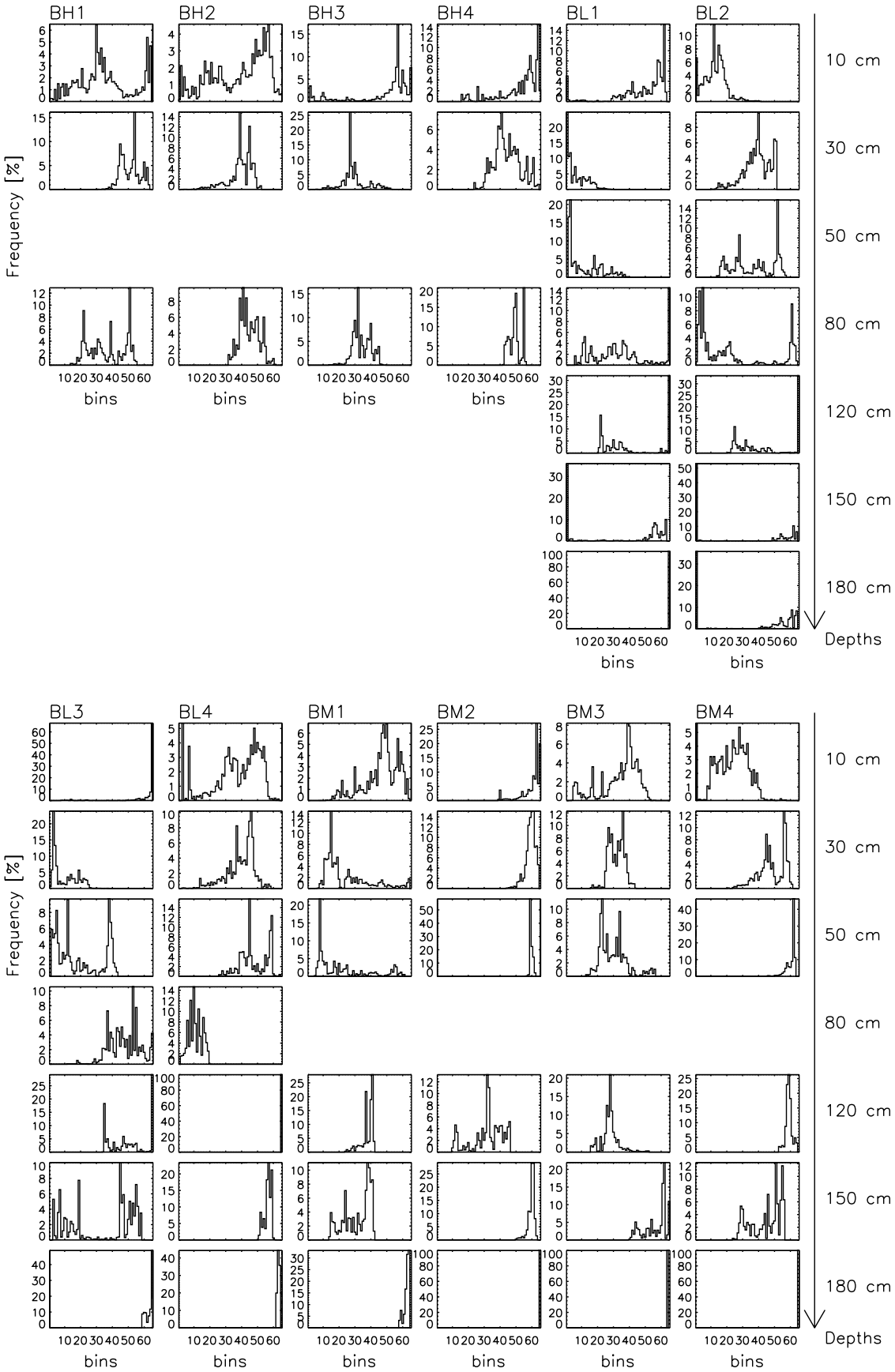

Figure 10. Talagrand diagrams for 64 ensemble members for all sensors in the B field at all depths over a period of 1 year (1 May 2001 to 1 May 2002) for perturbation of initial states and parameters.

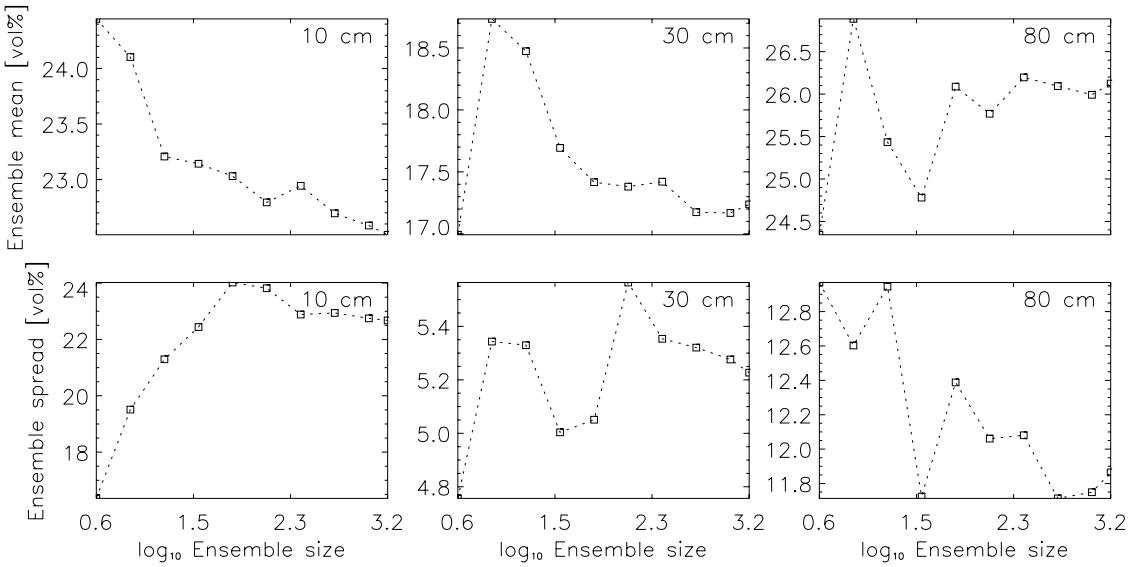

Figure 11. Evolution of the temporally averaged ensemble mean and spread for probe BH1 for increasing ensemble sizes. The ensembles were obtained by perturbation of initial states as well as parameters.

larger ensemble sizes the findings are similar, this shortcoming is often caused by model error, present at some depths. For probe BL2 for example, the Talagrand diagrams for the deepest layers show lack of variability in the ensembles (U shape). This is because the observations in these layers show evidence of lateral preferential flow, which causes a large difference in soil moisture before and after dry-out, which cannot be captured by the model. The J-shaped diagrams for the sensor BL4 at 120 cm and 180 cm indicate the presence of bias. It is remarkable that for H-profiles, better uniformity for all three layers is found than for profiles that are calibrated for more depths. This is in agreement with the finding that it is easier to find a best parameter set that yields good results over only three depths than calibration for six or seven depths at once.

6.5. Ensemble Statistics and Scores Versus Ensemble Size

[52] Several statistics and measures were studied as function of the ensemble size. The correlation, *RMSE*, and slope of the regression between the ensemble means and the control runs for individual soil layers were calculated for increasing ensemble sizes. Further, time-averaged ensemble mean, spread, kurtosis and skewness (see above time series) were calculated, and the measures $\frac{\langle\sqrt{ensk}\rangle}{\langle\sqrt{mse}\rangle}$ and $\frac{\langle ensk\rangle}{\langle ensp\rangle}$ were determined for varying ensemble sizes. The shapes of Talagrand diagrams were studied for their dependence on the ensemble size.

[53] We found that for most sensors the information from additional members is not worth the increase in computational effort from approximately 64 members on (e.g., Figure 11 in case of perturbation of initial conditions and parameters). For the different kinds of perturbations, similar evolutions of the statistics in function of the ensemble size are found. It should be remarked that an ensemble size of 64 is very low to determine the correct covariances between state errors for state estimation purposes: The amount of parameters is very high and it is very likely that 64 ensemble members will not be able to capture the covariance between changes of a parameter in one state variable with changes in another state variable. It is expected that the optimal ensemble size also depends on the chosen standard deviation in the ensemble generation, but this has not been explored; only realistic perturbations were simulated.

6.6. Measures of Goodness-of-Fit

[54] Results of the ensemble mean validation (3 October 2001 to 1 May 2002) for different kinds of ensemble perturbations against observations are summarized in Tables 7 and 8 for the *RMSE* of the probes in field B. For perturbation of initial states only, the *RMSE* values were almost identical to the values for validation of the control run and very little affected by the ensemble size (data not shown). Also for perturbation of forcings only, the *RMSE* values were marginally influenced by the ensemble size. The results for perturbation of initial conditions and parameters (Table 7) are almost equal to those for perturbation of parameters only. The ensemble mean model result yields for some profiles a better estimate of soil moisture than the control run, while for others the ensemble mean does worse, which is quite cumbersome, as we would have expected to improve results through the use of ensemble runs. However, the analyses in the above sections already showed the possibility of additional bias in the ensemble states by perturbing parameters. In general, there is only a small impact of the ensemble size on performance indices, with large ensemble sizes in general resulting in a slightly better performance. The variation of the values of the measures with the ensemble size is dependent on the sensor. Concerning the different types of perturbations, it is striking that with inclusion of forcing perturbation, the performance is in general slightly better than when only parameters and initial states are perturbed. This is possibly due to some compensation of errors.

7. Conclusions

[55] The subdivision of grid cells and patches in the CLM2.0 was explored for the generation of Monte Carlo simulations for use in calibration and ensemble generation.

A distributed multiobjective calibration was developed for the optimal estimation of parameters for 36 soil moisture profiles in space by a random search method. Several objective functions were defined and aggregated over the profiles. Through iterative sorting on different measures-of-fit, the best parameter set was obtained. Because the selection of optimal parameters depends on the definition of initial state variables, these were included in a least squares sense in one objective function that was used for calibration. This was basically a simple approach to weak constraint variational data assimilation.

[56] Because there is no indication that the resulting parameter and initial state values provide the ultimate best soil moisture solutions, and to better understand the forecast uncertainty, ensemble runs have been generated. Several methods commonly used in meteorology to assess the reliability of ensemble pdfs were applied to study ensembles of soil moisture generated by the CLM2.0.

[57] In the CLM2.0 and in all land surface models that are controlled by deterministic (atmospheric) forcings, the state does not evolve as freely as in, for example, atmospheric models because the dispersion is bounded. It was found that perturbation of the initial state variables only or forcings only does not suffice to describe the uncertainty of the state. The initial spread caused by initial state perturbation is dampened out and almost removed relatively quickly. Therefore the strong constraint approach is rejected for this study. The spread generated by perturbation of forcings only better explains the mse_i. Parameter uncertainty is of major importance and a weak constraint approach is needed. If in addition to parameter and initial state perturbation, the forcings are perturbed realistically, the spread does not change much, but the model performance enhances slightly. It can be concluded that realistic perturbation of parameters and forcings is most effective to generate ensemble members for land surface models. The optimal ensemble size may depend on the magnitude of ensemble perturbation. In this study, 64 members were found to be sufficient to represent the ensemble pdf.

[58] From the time series of ensemble pdfs and their moments, it is important to remember that rainfall events and extreme drought have a great impact on the shape of the distribution and cause strong nonnormalities. Further, it was found that the ensemble generated by random perturbation of the optimal parameters obtained through calibration does not always overlap with the theoretical random perturbation of the control run. Additional bias is sometimes introduced through ensemble generation.

[59] The results indicate that careful investigation of ensembles generated by random perturbation of parameters, initial states, and forcings is needed, before they can be used as an indication for forecast error. For state estimation, propagation of the mean and covariance as in the procedure for the ensemble Kalman filter is advised to minimize the errors that will inevitably be introduced by linearization of land surface models during crucial periods of precipitation or drought stress. However, one should carefully examine the ensemble mean behavior in advance, instead of assuming that it is always the best a priori estimate of the truth in an ensemble Kalman filter. Furthermore, the update equation will yield optimal results for the a posteriori estimate in the class of linear filters only, since the pdfs of the a priori state were clearly shown to deviate far from Gaussian. Improved estimates may be expected from nonlinear filters.

Table 7. Validation Measure of Goodness-of-Fit (*RMSE*, vol%) for Means of Different Ensemble Sizes, Generated by Perturbation of Initial States and Parameters, for All Sensors in Field B

	Ensemble Size										
Sensor	4	8	16	32	64	128	256	512	1024	1500	64[a]
BH1	3.26	2.31	2.35	2.66	2.50	2.49	2.43	2.49	2.50	2.46	3.11
BH2	5.45	4.79	2.97	2.90	2.76	2.80	2.75	2.75	2.74	2.75	2.91
BH3	1.97	2.00	2.47	2.40	2.41	2.34	2.41	2.28	2.34	2.34	2.25
BH4	5.82	7.19	8.55	9.35	8.02	7.34	7.02	6.84	6.84	6.95	3.30
BL1	8.51	8.34	8.31	8.42	8.64	8.73	8.69	8.75	8.85	8.86	9.14
BL2	9.55	9.46	9.76	9.05	8.78	8.67	8.71	8.60	8.66	8.69	9.13
BL3	6.12	6.76	6.19	6.35	6.41	6.25	6.34	6.39	6.45	6.46	6.65
BL4	7.32	6.99	6.62	6.45	6.39	6.27	6.34	6.42	6.38	6.41	4.35
BM1	5.17	5.05	5.02	4.90	4.96	4.74	4.67	4.72	4.67	4.66	4.36
BM2	11.19	10.80	10.77	10.99	10.92	10.91	11.06	11.12	11.05	11.09	9.33
BM3	3.09	3.78	4.05	3.99	4.34	4.44	4.46	4.47	4.43	4.45	4.59
BM4	7.54	7.39	7.46	7.70	7.41	7.45	7.35	7.35	7.35	7.29	6.48

[a]This column shows the *RMSE* (vol%), when 64 ensemble members are generated by perturbation of forcings only.

Table 8. Validation Measure of Goodness-of-Fit (*RMSE*, vol%) for Means of Different Ensemble Sizes, Generated by Perturbation of Initial States, Parameters, and Forcings, for All Sensors in Field B

	Ensemble Size									
Sensor	4	8	16	32	64	128	256	512	1024	1500
BH1	3.04	2.30	2.16	2.43	2.34	2.27	2.27	2.32	2.32	2.29
BH2	5.58	4.88	2.97	2.84	2.64	2.70	2.62	2.64	2.62	2.63
BH3	2.00	1.95	2.49	2.45	2.46	2.41	2.45	2.31	2.37	2.37
BH4	5.77	7.02	8.40	9.21	7.89	7.22	6.91	6.72	6.72	6.83
BL1	8.56	8.37	8.34	8.47	8.70	8.80	8.76	8.82	8.93	8.93
BL2	9.88	9.75	10.07	9.33	9.05	8.93	8.97	8.87	8.93	8.95
BL3	6.01	6.68	6.14	6.32	6.36	6.23	6.33	6.38	6.45	6.46
BL4	7.32	6.95	6.52	6.34	6.28	6.17	6.24	6.32	6.27	6.30
BM1	5.10	4.98	4.96	4.83	4.76	4.59	4.54	4.59	4.56	4.57
BM2	11.00	10.61	10.59	10.84	10.79	10.78	10.93	10.99	10.92	10.96
BM3	3.02	3.70	4.00	3.93	4.29	4.34	4.36	4.37	4.33	4.35
BM4	7.30	7.10	7.18	7.45	7.16	7.11	7.02	7.08	7.10	7.04

[60] **Acknowledgments.** The authors thank the Beltsville Agricultural Research Center (BARC)–Agricultural Research Service (ARS) of the USDA for providing the data. The Hydrological Sciences Branch (HSB) of NASA/GSFC is thanked for hosting the first author during part of the research. The research is supported by a PhD-fund of the Bijzonder Onderzoeksfonds (BOF) of Ghent University and partly by STEREO project SR/00/01 of the Belgian Science Policy (BELSPO). The authors thank the reviewers for their constructive comments.

References

Anderson, J. (1996), A method for producing and evaluating probabilistic forecasts from ensemble model integrations, *J. Clim.*, *9*, 1518–1530.

Anderson, J. L., and S. L. Anderson (1999), A Monte Carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts, *Mon. Weather Rev.*, *127*, 2741–2758.

Bastidas, L. A., H. V. Gupta, S. Sorooshian, W. J. Shuttleworth, and Z. L. Yang (1999), Sensitivity analysis of a land surface scheme using multicriteria methods, *J. Geophys. Res.*, *104*(D16), 19,481–19,490.

Beven, K. (1993), Prophecy, reality and uncertainty in distributed hydrological modelling, *Adv. Water Resour.*, *16*, 42–51.

Beven, K., and J. Freer (2001), Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology, *J. Hydrol.*, *249*, 11–29.

Bonan, G. (1996), A land surface model (LSM version 1.0) for ecological, hydrological, and atmospheric studies, technical report, Natl. Cent. for Atmos. Res., Boulder, Colo.

Boyle, D. P., H. V. Gupta, and S. Sorooshian (2000), Toward improved calibration of hydrologic models: Combining the strengths of manual and automatic methods, *Water Resour. Res.*, *36*(12), 3663–3674.

Boyle, D. P., H. V. Gupta, S. Sorooshian, V. Koren, Z. Zhang, and M. Smith (2001), Toward improved streamflow forecasts: Value of semidistributed modeling, *Water Resour. Res.*, *37*(11), 2749–2759.

Brankovic, C., T. N. Palmer, F. Molteni, S. Tibaldi, and U. Cubasch (1990), Extended-range predictions with ECMWF models time-lagged ensemble forecasting, *Q. J. R. Meteorol. Soc.*, *116*, 867–912.

Butts, M. B., J. T. Payne, M. Kristensen, and H. Madsen (2004), An evaluation of the impact of model structure on hydrological modelling uncertainty for streamflow simulation, *J. Hydrol.*, *298*, 242–266.

Carpenter, T. M., and K. P. Georgakakos (2004), Impacts of parametric and radar rainfall uncertainty on the ensemble streamflow simulations of a distributed hydrologic model, *J. Hydrol.*, *298*, 202–221.

Clapp, R., and G. Hornberger (1978), Empirical equations for some soil hydraulic properties, *Water Resour. Res.*, *14*(4), 601–604.

Cramér, H. (1946), *Mathematical Methods of Statistics*, Princeton Univ. Press, Princeton, N. J.

Crow, W. (2003), Correcting land surface model predictions for the impact of temporally sparse rainfall rate measurements using an ensemble Kalman filter and surface brightness temperature observations, *J. Hydrometeorol.*, *4*, 960–973.

Crow, W. T., and E. F. Wood (2003), The assimilation of remotely sensed soil brightness temperature imagery into a land surface model using ensemble Kalman filtering: A case study based on ESTAR measurements during SGP97, *Adv. Water Resour.*, *26*, 137–149.

Dai, Y., and X. Zeng (1996), A land surface model (IAP94) for climate studies: Part I. Formulation and validation in off-line experiments, *Adv. Atmos. Sci.*, *14*, 433–460.

De Lannoy, G. J. M., N. E. C. Verhoest, and F. P. De Troch (2005), Characteristics of rainstorms over a temperate region derived from multiple time series of weather radar images, *J. Hydrol.*, *307*, 126–144.

Dickinson, R. E., A. Henderson-Sellers, and P. Kennedy (1993), Biosphere-Atmosphere Transfer Scheme (BATS) version 1e as coupled to the NCAR Community Climate Model, technical report, Natl. Cent. for Atmos. Res., Boulder, Colo.

Evensen, G. (1994), Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics, *J. Geophys. Res.*, *99*(C5), 10,143–10,162.

Gao, X., S. Sorooshian, and H. V. Gupta (1996), Sensitivity analysis of the biosphere-atmosphere transfer scheme, *J. Geophys. Res.*, *101*(D3), 7279–7289.

Georgakakos, K., D.-J. Seo, H. V. Gupta, J. Schaake, and M. Butts (2004), Towards the characterization of streamflow simulation uncertainty through multimodel ensembles, *J. Hydrol.*, *298*, 222–241.

Gupta, H. V., S. Sorooshian, and P. O. Yapo (1998), Toward improved calibration of hydrological models: Multiple and noncommensurable measures of information, *Water Resour. Res.*, *34*(4), 751–763.

Hamill, T. M. (2001), Interpretation of rank histograms for verifying ensemble forecasts, *Mon. Weather Rev.*, *129*, 550–560.

Houser, P. R., H. V. Gupta, J. Shuttleworth, and J. S. Famiglietti (2001), Multiobjective calibration and sensitivity of a distributed land surface water and energy balance model, *Water Resour. Res.*, *106*(D24), 33,421–33,433.

Kalman, R. E. (1960), A new approach to linear filtering and prediction problems, *J. Basic Eng.*, *82*(D), 35–45.

Krzysztofowicz, R. (2001), The case for probabilistic forecasting in hydrology, *J. Hydrol.*, *249*, 2–9.

Lee, K.-H., and E. N. Anagnostou (2004), Investigation of the nonlinear hydrologic response to precipitation forcing in physically based land surface modeling, *Can. J. Remote Sens.*, *30*(5), 706–716.

Margulis, S. A., D. McLaughlin, D. Entekhabi, and S. Dunne (2002), Land data assimilation of soil moisture using measurements from the Southern Great Plains 1997 Field Experiment, *Water Resour. Res.*, *38*(12), 1299, doi:10.1029/2001WR001114.

Miller, R. N., E. Carter, and S. T. Blue (1999), Data assimilation into nonlinear stochastic models, *Tellus, Ser. A*, *51*, 167–194.

Reichle, R. H., D. B. McLaughlin, and D. Entekhabi (2002a), Hydrologic data assimilation with the Ensemble Kalman filter, *Mon. Weather Rev.*, *120*, 103–114.

Reichle, R. H., J. P. Walker, P. R. Houser, and R. D. Koster (2002b), Extended vs. ensemble Kalman filtering for land data assimilation, *J. Hydrometeorol.*, *3*, 728–740.

Starr, J., and I. C. Paltineanu (2002), Methods for measurement of soil water content: Capacitance devices, in *Methods of Soil Analysis: Part 4: Physical Methods*, edited by J. Dane and G. Topp, pp. 463–474, Soil Sci. Soc. of Am., Madison, Wis.

Stephenson, D. B., and F. J. Doblas-Reyes (2000), Statistical methods for interpreting Monte Carlo ensemble forecasts, *Tellus, Ser. A*, *52*, 300–322.

Talagrand, O., R. Vautard, and B. Strauss (1997), Evaluation of probabilistic prediction systems, technical report, Eur. Cent. for Medium-Range Weather Forecast., Reading, UK.

Toth, Z., O. Talagrand, G. Candille, and Y. Zhu (2003), Probability and ensemble forecasts, in *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, edited by I. Jolliffe and D. Stephenson, chap. 7, pp. 137–164, John Wiley, Hoboken, N. J.

Xia, Y., Z.-L. Yang, P. Stoffa, and M. Sen (2005), Using different hydrological variables to assess the impact of atmospheric forcing errors on optimization and uncertainty analysis of the CHASM surface model, *J. Geophys. Res.*, *110*, D01101, doi:10.1029/2004JD005130.

Yapo, P., H. Gupta, and S. Sorooshian (1996), Automatic calibration of conceptual rainfall-runoff models: Sensitivity to calibration data, *J. Hydrol.*, *181*, 23–48.

Zeng, X. (2003), The Common Land Model experience, *Global Change Newsl.*, *55*, 19–20.

G. J. M. De Lannoy, V. R. N. Pauwels, and N. E. C. Verhoest, Laboratory of Hydrology and Water Management, Ghent University, Coupure links 653, B-9000 Ghent, Belgium. (gabrielle.delannoy@ugent.be)

P. R. Houser, George Mason University and Center for Research on Environment and Water, 4041 Powder Mill Road, Suite 302, Calverton, MD 20705-3106, USA.